УДК 519.766

1.1. О консенсусе языковых моделей искусственного интеллекта

ЕРЕШКО Ф. И.
профессор, д. т. н., Вычислительный центр им. А. А. Дородницына Российской академии наук Федерального исследовательского центра «Информатика и управление» РАН,
САРАЕВ В.Н.,
к.т.н., Международный научно-исследовательский института, АО “РК сложные системы” Фонда Росконгресс;
ТКАЧЕНКО А. В.
Аспирант, Институт проблем управления им. В. А. Трапезникова РАН

Рассмотрены языковые модели, модели трансформера искусственного интеллекта сопряженные с процедурами теории игр. В предложенной Схеме решения конкретных задач с помощью искусственного интеллекта на основе игры консенсусный трансформер должен имитировать действия физического лица. Трансформер формирует стратегию на естественном языке.

ВВЕДЕНИЕ.

Президент России В.В. Путин 10 апреля 2026 г. провёл в Кремле совещание по вопросам развития технологий искусственного интеллекта. В ходе совещания он отметил, что “ искусственный интеллект наряду с цифровыми платформами и автономными системами формирует принципиально иной облик экономики, общественных отношений, социальной сферы, образования, здравоохранения, логистики и промышленности, обороны и безопасности…Что хочу особо подчеркнуть: языковые модели – это базовая, сквозная технология, которая является основой для суверенного развития всех сфер. И только при наличии собственных моделей мы сможем уверенно двигаться вперёд, гарантировать безопасность и обороноспособность, что особенно важно, оставаться на переднем крае научной и инженерной мысли, тем более что наша страна в числе немногих государств обладает в этой сфере уникальными компетенциями.”

Президент России В.В. Путин поручил Правительству совместно с главами регионов сформировать “Национальный план внедрения искусственного интеллекта на уровне всей страны с учётом задач отраслей и субъектов Федерации.” Вопросам формирования Национального плана через призму динамики развития языковых моделей посвящена эта статья.

АНАЛИЗ СЛОЖНЫХ СИСТЕМ

Исследования имитационных игр при анализе сложных систем столкнулись с внутренними противоречиями их составляющих с одной стороны неопределенностью намерений, с другой многокритериальностью устремлений. Диалектический процесс тезиса, антитезиса и синтеза [1] в школе академика Моисеева Н. Н. [2] был осуществлен в виде следующего имитационного подхода: результаты реализации реального процесса, построенного на основе математической модели, предлагаются реальным участникам имитационной игры, которые затем осуществляют на их основе принятие решения. Например, в Игре “ВОЙНА И МИР”. МЕЖСТРАНОВОЕ ПРОТИВОБОРСТВО трех стран СССР, США и третьего мира имитация использовалась как поддержка принятия решений, играть должны были люди. Разработчиком игры был член-корреспондент РАН Павловский Ю.Н. [3].

“У стран была собственная экономика. Экономика состояла из двух секторов мирного и военного. Военный сектор выпускал вооружения, которыми можно было впоследствии воевать. Мирный сектор выпускал мирную продукцию, которую в дальнейшем можно было инвестировать, либо в мирный сектор, либо в военный сектор. Неинвестированный мирный продукт оставался в запасе. Его можно было хранить, инвестировать где-то в будущем, или, например, торговать им с другими странами. Можно было торговать, назначать цены, договариваться о чем угодно, можно было обманывать и даже воевать.” (см. рис. 1) [4].

В работе [5] было предложено, что роль одного из игроков исполняет трансформер (устройство трансформации информации – вычислительный комплекс). Выражение трансформер может употребляться в разных смыслах: как модель, программа, стиль архитектуры [6]. Трансформер (transformer) в среде языковых моделей искусственного интеллекта– это нейросетевая архитектура на основе распараллеливаемости процедур моделей внимания и полносвязанных слоев [7], которая лежит в основе большинства современных больших языковых моделей (LLM).

ЯЗЫКОВЫЕ МОДЕЛИ

“Языковая модель — это распределение вероятностей, определенное для последовательности слов (предложения или абзаца)” [8]. В основе механизма моделирования текстов на естественном языке, основанного на теории вероятностей, статистике и машинном обучении, лежат языковые модели. Моделирование естественного языка основано или на теории вероятности или теории формального языка. “Цепь Маркова”, базисом которой являются вероятности перехода между состояниями цепи, предложенная в 1906 г. академиком Императорской Санкт-Петербургской академии наук А.А. Марковым, является генезисом исследования языковых моделей. [9]. В

цепи Маркова вероятность наступления каждого события зависит только от состояния, достигнутого в предыдущем событии. Модель была протестирована Марковым в 1913 г. на романе в стихах Александра Пушкина "Евгений Онегин". Пример языковой модель в виде цепи Маркова выглядит следующим образом. Пусть $\omega_1, \omega_2, \ldots\omega_n$ –это последовательность слов и под обозначением $P(\omega_1|\omega_0)=P(\omega_1)$ будем понимать просто $P(\omega_1)$. Тогда вероятность встречаемости последовательности слов может быть

$$P(\omega_1, \omega_2, \ldots\omega_n)= \prod_{i=1}^{n} P(\omega_i| \omega_1, \omega_2, \ldots\omega_{i-1}) \quad (1).$$

Например, для $n=5$,

$P(\omega_1, \omega_2, \omega_3, \omega_4,\omega_5)= P(\omega_1)* P(\omega_2|\omega_1)* P(\omega_3|\omega_1, \omega_2)* P(\omega_4|\omega_1, \omega_2, \omega_3)* P(\omega_5|\omega_1, \omega_2, \omega_3, \omega_4)$ (2).

Различные типы языковых моделей используют различные методы для вычисления условных вероятностей вида $P(\omega_i|\omega_1, \omega_2, \ldots\omega_{i-1})$. Базовая языковая модель – это модель n-грамма, которая предполагает, что слово в каждой позиции зависит только от слов в $n-1$ предыдущих.

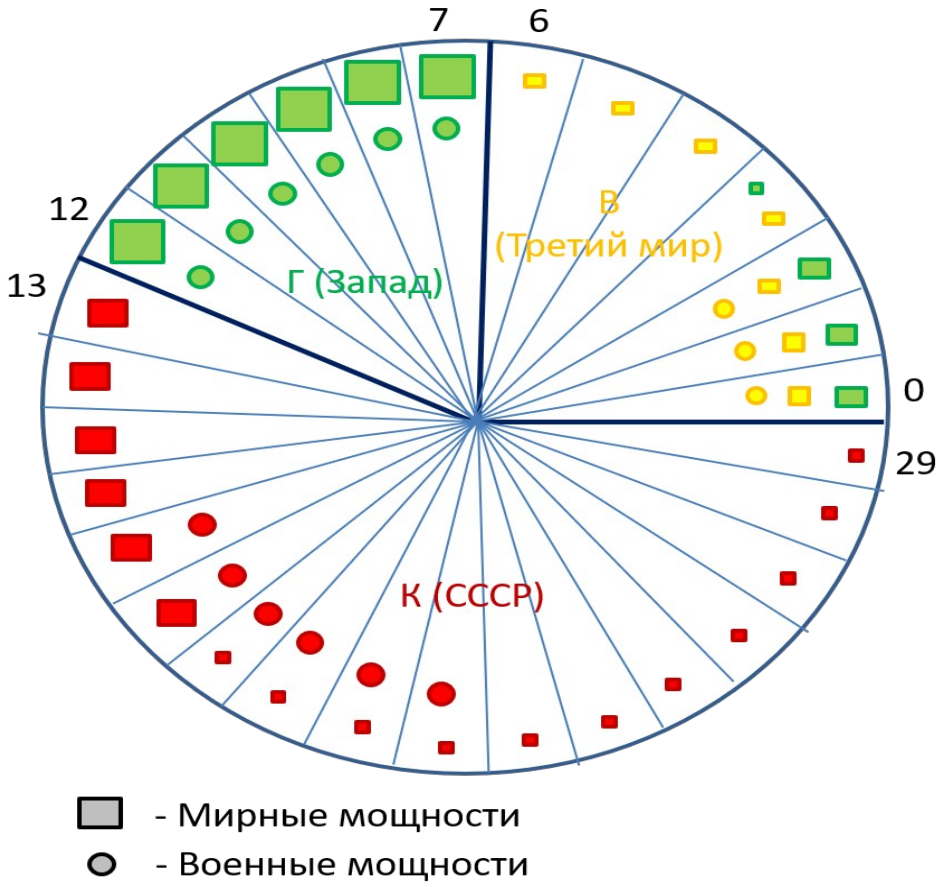

Рис. 1 Начальное состояние стран.

Американский ученый Клод Шеннон в 1948 году разработал математическую теорию связи, в которой ввел понятие энтропии, кросс-энтропии и рассмотрел свойства n-граммовой модели [10]. Если энтропия — это неопределенность одного распределения вероятностей, то кросс-энтропия (перекрёстная энтропия) это неопределенность одного распределения вероятностей по отношению к другому распределению вероятностей. Чем меньше перекрёстная энтропия, тем более точно языковая модель будет предсказывать последовательность слов.

Другим направлением формализации языковой модели, кроме вероятностного подхода, является теория формального языка, на основе иерархии грамматик Ноама Хомского (1956 г.) [11]. Следует отметить, что грамматики с конечным состоянием (конечной цепи Маркова или n-граммовой модели), также имеют ограничения при описании естественных языков.

ТРАНСФОМЕР

Дальнейшие процедуры формализации языковой модели привели к модели трансформера [7], которая избегает рекуррентности формируя механизм внимания для построения глобальных зависимостей между входной и выходной последовательностями. Функцию внимания можно описать как сопоставление запроса (query) и набора пар ключ-значение (key-value) с выходными данными, где запрос, ключи, значения и выходные данные являются векторами. Результат вычисляется как взвешенная сумма значений, где вес, присвоенный каждому значению, вычисляется функцией совместимости запроса и соответствующего ключа.

В игре трансформер должен имитировать действия физического лица. Для участия в игре трансформер должен пройти обучение в серии игр, и научиться вырабатывать стратегию принятия решения, описываемую естественным языком (NLM), т. е. он должен принимать решения как физическое лицо. “В определенном смысле, это соответствие тесту Тьюринга” [5].

Действия трансформера аналогичны ниже представленной Схеме преобразования данных в машинном переводе, на интервале развития процесса ИГРЫ [t,t+1]:

1.В момент t КОДИРОВЩИК преобразует данные на языке ИГРЫ во входные данные последующей языковой модели на естественном языке.

2.На интервале ИГРЫ [t,t+1] входные данные на естественном языке ЯЗЫКОВАЯ МОДЕЛЬ (Chat GPT, Deep Seek) трансформирует в выходные данные языковой модели на естественном языке.

*3.*ДеКОДИРОВЩИК в момент t+1 преобразует выходные данные языковой модели в слова предложения на машинном языке для получения данных на языке описания игрового процесса в продолжение игры.

Таким образом, Трансформер формирует стратегию на естественном языке.

Алгоритм работы трансформера в Схеме преобразования данных в машинном переводе выглядит следующим образом [6,7]:

- $S = (g_1,\ldots, g_n)$ – слова предложений на входном языке
  ↓ обучаемая или пред-обученная векторизация слов
- $X= (x_1,\ldots,x_n)$ – векторы слов входного предложения
  ↓ трансформер-кодировщик
- $Z = (z_1,\ldots,z_n)$ – контекстные векторы слов
  ↓ трансформер-декодировщик, похож на кодировщика
- $Y – (y_1,\ldots,y_m)$ – векторы слов выходного предложения
  ↓ генерация слов из построенной языковой модели
- $\check{S} =(\check{g}_1,\ldots,\check{g}_m)$ слова предложения на выходном языке.

Для игрового воспроизведения искусственных нейронных сетей трансформер выглядит как нейросетевая архитектура, представленная на основе модели игры, которая рассмотрена в [12]. Были исследованы две модели обработки информации в виде полносвязной искусственной нейронной сети. В первом случае функции активации задаются исследователем. Во втором являются выбором искусственного нейрона-его стратегией.

Большие языковые модели (LLM), в том числе и построенные на основе архитектуры трансформера, дают разные ответы на один и тот же вопрос, представленный в разной форме. Генеративный вопрос (“Какая столица России?” - Москва), который является открытым дает один ответ, дискриминативный вопрос (“Москва- столица России?”), который подразумевает необходимость выбора между вариантами, часто дает другой ответ.

Генеративный подход утверждает, что реальность субъективна изначальна, она создается нами в процессе нашего восприятия. Утверждение Альфреда Корбжинского “Карта не равна территории” [13] показывает, что восприятие реальности может не совпадать с реальностью. Уточнение Ричарда Блендера “Карта не равна карте” сообщает, что о самой реальности мы ничего не можем знать, у нас есть только наше восприятие. Наше восприятие отличается от восприятия другого человека [14]. Генеративность (generatio-лат. – рождение, производство) – часть творчества, процесс рождения нового. Задача генеративного моделирования – это создание реалистического изображения, например собаки или кошки. Задача дискриминативного моделирования — это классификация изображений, например собак или кошек. Генеративный вопрос – это вопрос, побуждающий к созданию ответа, не содержащегося в вопросе.

Процедура согласования двух языковых моделей – генератора, который обрабатывает генеративные вопросы, и дискриминатора, который обрабатывает дискриминативные вопросы, в работе [15] представлена в виде игры в консенсус, в которой заложена идея разговора между двумя людьми, в котором слушатель стремиться понять, что пытается сказать говорящий.

Генератор получает вопрос без ответов, а дискриминатор получает несколько возможных ответов от человека, списка или поиска, выполняемого самой языковой моделью. Но у генератора список возможных ответов содержится (неявным образом) в самой модели.

В начале процедуры согласования, в результате случайного подбрасывания монеты, определяют корректным или некорректным должен быть ответ генератора. Если выпадает орел – генератор пытается ответить корректно и отправляет исходный вопрос вместе с ответом дискриминатору. Если дискриминатор определяет, что генератор намеренно отправил корректный ответ, каждый из них получает по одному очку в качестве своего рода стимула. Если выпадает решка, генератор отправляет дискриминатору то, что считает некорректным ответом. Если дискриминатор определяет, что ему намеренно отправили некорректный ответ, то они оба также получают по одному очку в качестве стимула.

Генератор и дискриминатор начинают процедуру согласования с некоторых начальных “убеждений”, которые имеют разные формы распределения вероятности, связанные с различными вариантами выбора (не явных для генератора и явных – для дискриминатора). У дискриминатора может быть другая форма распределения

вероятности, которая формируется, например, из интернета. Языковые модели назначают разные вероятности разным вариантам ответа. При слишком большом отклонении от первоначальных убеждений в процессе процедуры согласования генератор и дискриминатор могут быть наказаны. Такой алгоритм требует от игроков включать свои знания о мире, снова взятые из интернета в свои ответы, что возможно, могло бы улучшать точность модели.

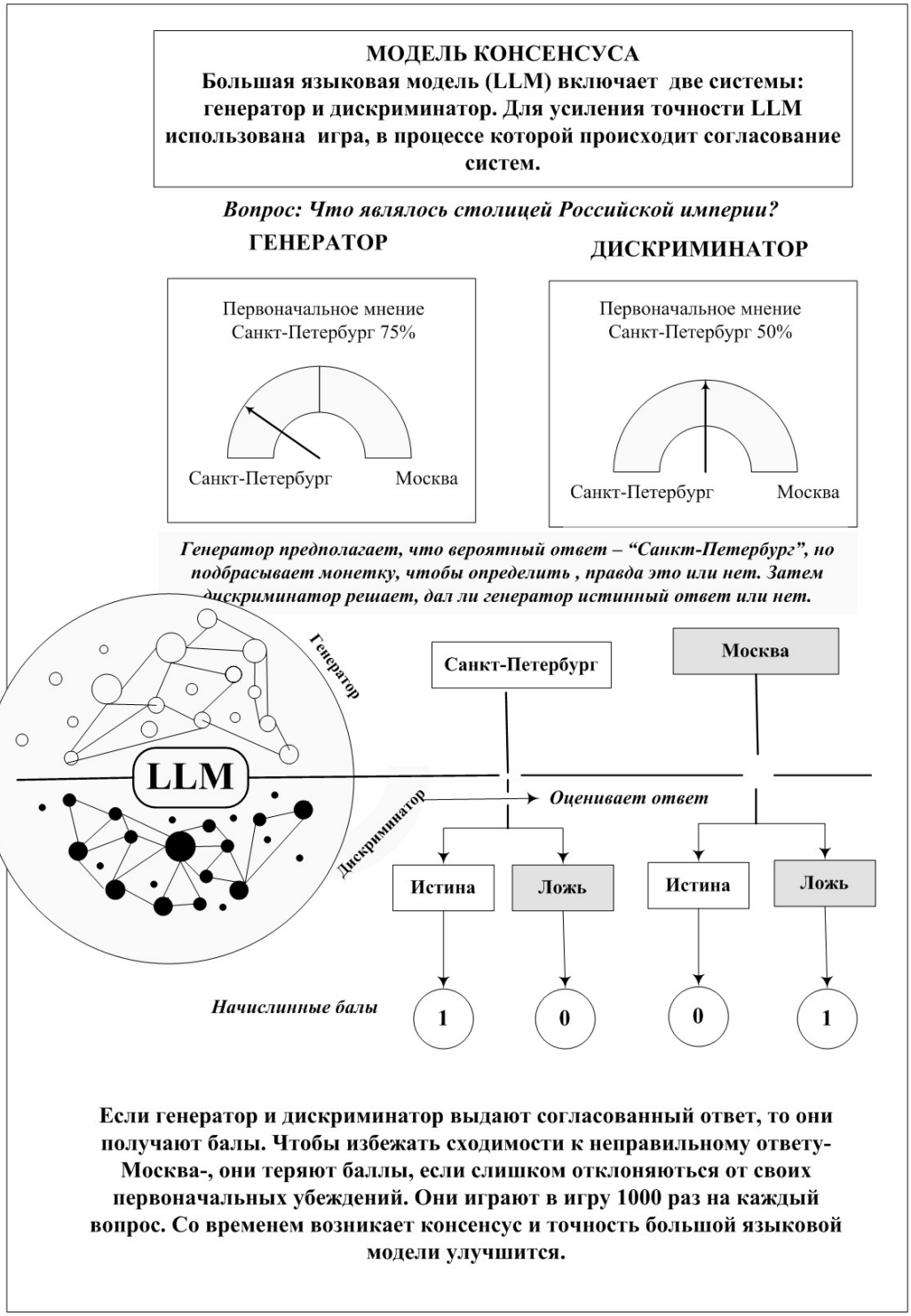

Рис.2. Источник [17].

В процессе согласования языковая модель (LM) сопоставляет входные строки *x* с выходными строками *y* в соответствии с некоторым вероятностным распределением $P_{LM}(y|x)$. Например (см. рис.2) в ответе на вопрос *X* (столица Российской империи?) и набора возможных ответов *Y* (Санкт-Петербург, Москва,…), которые имеют

разное вероятностное распределение $P_{LM}(\cdot|x)$. В стратегии этот набор кандидатов на ответ Y можно использовать (по крайней мере) двумя способами:

1.Генератор получает входные данные в виде:

а) запроса X (опционально),

б) набора кандидатов на ответ Y,

в) запроса на естественном языке, указывающего нужно ли выдать корректный или некорректный ответ.

В этом случае языковую модель LM можно рассматривать с вероятностным распределением $P_{LM}(y|x, correct)$, где токен “correct” означает тот факт, что модели было предложено сгенерировать корректный ответ.

2.Дискриминатор получает входные данные в виде:

а) запроса X (поставленного перед генератором),

б) набора кандидатов на ответ $y \epsilon Y$,

в) ответ генератору,

г) запроса (к самому дискриминатору), определить является ли ответ генератора корректным или не корректным.

В этом случае языковая модель (LM) выступает как инструмент моделирования с вероятностным распределением $P_{LM}(v|x,y)$, где $v\epsilon$(correct, incorrect).

Авторы статьи [15] утверждают, что эти два подхода концептуально эквиваленты, но современные языковые модели могут давать очень разные ответы, когда их спрашивают разными способами. Ответы, полученные в ответ на запрос в обобщенном виде с высокой вероятностью, могут на самом деле оказаться как верными, так и не верными.

Исследования, посвященные языковым моделям, показали [15], что решить эту проблему можно следующими способами:

- Методы создания ансамблей через простое объединение генеративных и дискриминационных оценок. Недостаток-оценки не проходят процедуру согласования.
- Методы обсуждения, в которых процедура согласования осуществляется в стратегии, например путем повторного запроса конкурирующих входных данных и инструкции по созданию текстового обоснования наилучшего варианта. Недостаток – большие вычислительные затраты.

Авторы [15], утверждают, что эффективная процедура для достижения консенсуса между конкурирующими вариантами ответа языковых моделей должна обладать двумя ключевыми свойствами:

1.процедуры генеративных и дискриминационных оценок должны определять какие кандидаты ответов являются правильными.

2.варианты ответов должны быть не произвольными, а максимально приближенными к генеративным и дискриминационным оценкам.

Формальной основой этих требований в теории игр является равновесие по Нэшу [18]. Как показывает равновесие по Нэшу такая некооперативная игра является самоподдерживающейся. Игрок А применяет стратегию, которая является его лучшим ответом на стратегию, предложенную игроком Б, который соблюдает правила игры. В результате такой стратегии возникает равновесие Нэша.

Процедуру декодирования можно представить на языке теории игр и вычислить стратегии равновесия в этой игре для получения консенсуса [16]. Игра (для N игроков) определяется пространством состояний S, пространством стратегий A, (детерминированной) функцией перехода $T: S \times A^N \rightarrow S$ и набором функций вознаграждения U_i. Поведение каждого игрока в игре моделируется как стратегия $f_i: S \rightarrow \infty(A)$ (распределение по стратегиям в заданных состояниях). В каждом раунде игры каждый игрок наблюдает (возможно, неполное) представление о текущем состоянии. Один или несколько игроков выбирают действия $a^t i \in u_i(\cdot|s^t_i)$, затем каждый игрок получает награду $u^t_i (s^t, a^t = a^t_1, \ldots, a^t_n)$, и игра переходит в новое состояние $s^{t+1} = T(s^t, a^t)$. Каждый игрок i стремится максимизировать ожидаемое вознаграждение u_i, и оптимальные стратегии для этого могут зависеть от следующих стратегий $f_{-i} = \{f_1, \ldots, f_{i-1}, f_{i+1}, \ldots..f_n\}$ других игроков. Игра на основе консенсуса проводится на игровом дереве (см. рис. 3). В начале игры определяется параметр корректности подбрасыванием монеты $v\epsilon$ {correct, incorrect}.

Параметр корректности учитывается только генератором и определяет должен ли генератор генерировать корректные или некорректные ответы. При соблюдении этого параметра ГЕНЕРАТОР выдают строку на естественном языке из фиксированного набора вариантов. Затем эти строки анализирует ДИСКРИМИНАТОР, который пытается определить значение параметра корректности v, выбирая в качестве ответа один из вариантов {correct, incorrect}. Оба игрока получают выигрыш в размере 1, если ДИСКРИМИНАТОР правильно определил значение параметра корректности, в противном случае их выигрыш равен нулю.

Ожидаемые выплаты по определению равны:

$$U_G(f_G,f_D) := 1/2 \sum_{v\epsilon \{\text{correct, incorrect}\}} \sum_{y\epsilon v} f_G(y|x,v)\, f_D(v|x,y) \quad (3),$$

$$U_D(f_G,f_D) := 1/2 \sum_{v\epsilon \{\text{correct, incorrect}\}} \sum_{y\epsilon v} f_G(y|x,v)\, f_D(v|x,y) \quad (4).$$

Эффективной стратегией согласно теории игр [18], является равновесие по Нэшу на основе пары стратегий (одна для ГЕНЕРАТОРА, другая для ДИСКРИМИНАТОРА), каждая из которых является оптимальной для одного игрока. То есть стратегия каждого игрока максимизирует его ожидания, учитывая стратегию другого игрока. При равновесии по Нэшу ни у одного игрока нет стимула в одностороннем порядке вести себя каким-либо иным

образом. Консенсусная игра допускает множество равновесий по Нэшу, несовместимых со здравым смыслом. Чтобы избежать не приемлемых равновесий в работе [15] был введен параметр регуляризации в полезность игроков так чтобы и ГЕНЕРАТОР, и ДИСКРИМИНАТОР получали штрафные санкции за выбор стратегий, которые далеки от некоторой пары исходных стратегий $f_G(1)$ и $f_D(1)$.

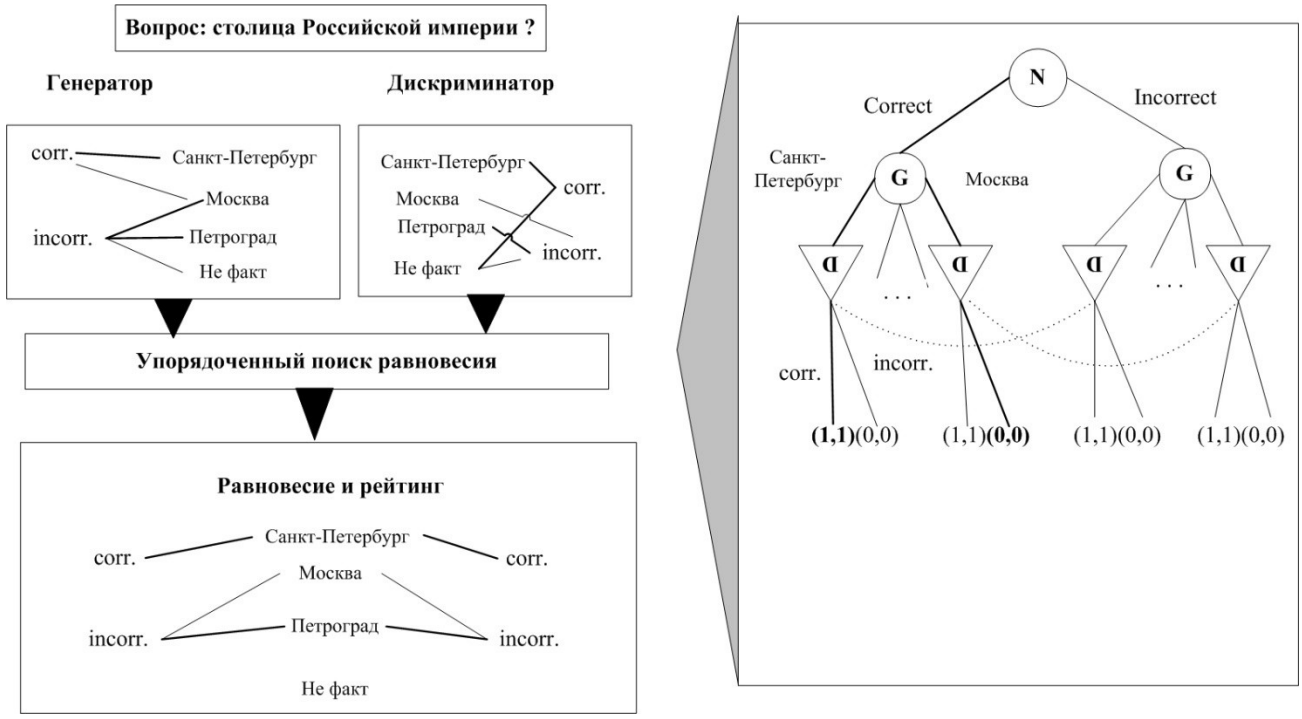

Рис.3: (слева) Различные запросы LM не позволяют прийти к единому мнению относительно ответа на фактический вопрос. Согласовывая прогнозы между генеративными и дискриминативными запросами LM с помощью игры КОНСЕНСУСА, получаем точный прогноз. (Справа) Структура КОНСЕНСУСНОЙ ИГРЫ, двухуровневой последовательной сигнальной игры с неполной информацией. Сначала (N) определяется параметр корректности. ГЕНЕРАТОР (G), настроенный на этот параметр, генерирует строку естественного языка из набора кандидатов. ДИСКРИМИНАТОР (D) наблюдает только за этой строкой и должен предсказать параметр корректности, выбранный системой. Если ДИСКРИМИНАТОР правильно определяет этот параметр, то оба игрока получают вознаграждение в размере 1. Пунктирная линия соединяет узлы, которые не различимы ДИСКРИМИНАТОРОМ, так как ДИСКРИМИНАТОР не учитывает параметр корректности. Вычисляя стратегии упорядоченного равновесия для этой игры, получают прогнозы, которые отражают консенсус между ГЕНЕРАТОРОМ и ДИСКРИМИНАТОРОМ. Источник [15].

Параметр регуляризации введен в функцию полезности, которую пытаются оптимизировать и ГЕНЕРАТОР, и ДИСКРИМИНАТОР. Вместо простой выплаты 0-1, определяемой по параметру корректности, происходит оптимизация функций, описываемых уравнениями 5-6.

$$U_G(f_G,f_D) := -\lambda_G \cdot D_{KL}[f_G(\cdot|x,v) \,||\, f_G^1(\cdot|x,v)] + 1/2 \sum_v \sum_{y \epsilon v} f_G(y|x,v) \cdot f_D(v|x,y) \quad (5),$$

$$U_D(f_G,f_D) := -\lambda_D \cdot D_{KL}[f_D(\cdot|x,y) \,||\, f_D^1(\cdot|x,y) + 1/2 \sum_v \sum_{y \epsilon v} f_G(y|x,v)\, f_D(v|x,y) \quad (6),$$

где λ_G и λ_D параметры регуляризации;

$D_{KL}(P|Q) = \sum P(x) \cdot log\ P(x)/Q(x)$ дивергенция Кульбаха-Лейбнера – мера того, насколько аппроксимирующее распределение вероятности Q отличается от истинного распределения вероятности P.

Следует отметить, что начальные стратегии $f_G(1)$ и $f_D(1)$ могут быть получены с помощью языковой модели, в которой задана первоначальная строка x для получения контекстных прогнозов (например, ответов на вопросы). Использование этой возможности в начальных стратегиях ГЕНЕРАТОРА и ДИСКРИМИНАТОР приводят равновесие Нэша в игре в направлении усиления консенсуса.

Итак, в результате выше рассмотренного можно сформулировать сопряжение в модели ИГРЫ консенсусного трансформера, построенного на основании теории игр, формирующего подобие русской матрешки, в которой существует игра в ИГРЕ.

КОНСЕНСУСНЫЙ ТРАНСФОРМЕР

Консенсусный трансформер формирует Схему преобразования данных в машинном переводе следующим образом:

- $S = (g_1,\ldots, g_n)$ – слова предложений на входном языке
  ↓ обучаемая или пред-обученная векторизация слов
- $X= (x_1,\ldots,x_n)$ **v** $Q=(q_1,\ldots q_n)$– векторы слов входного предложения
  ↓ трансформер-криптограф **v** трансформер-кодировщик
- $Z = (z_1,\ldots,z_n)$ – контекстные векторы слов

↓ трансформер- декриптограф**v**декодировщик, похож на криптографа и кодировщика

- $Y – (y_1,\ldots,y_m)$ – векторы слов выходного предложения
  ↓ генерация слов из построенной языковой модели
- $\check{S} =(\check{g}_1,\ldots,\check{g}_m)$ слова предложения на выходном языке.

Использование консенсусного трансформера преобразует процесс ИГРЫ на интервале [t,t+1] следующим образом:

1.В момент t КРИПТОГРАФ, сопряженный с КОДИРОВЩИКОМ, преобразует данные на языке ИГРЫ во входные данные последующей языковой модели на естественном языке.

2.На интервале ИГРЫ [t,t+1] входные данные на естественном языке ЯЗЫКОВАЯ МОДЕЛЬ (Chat GPT, Deep Seek) трансформирует в выходные данные языковой модели на естественном языке.

*3.*ДеКРИПТОГРАФ, сопряженный с ДеКОДИРОВЩИКом, в момент t+1 преобразует выходные данные языковой модели в слова предложения на машинном языке для получения данных на языке описания игрового процесса в продолжение игры.

МЕДИЦИНСКИЙ АССИСТЕНТ

Предложенный подход по формированию консенсусного трансформера позволяет преобразовать патентно-техническое решение ПАО Сбербанка интеллектуальный медицинский ассистент [19] следующим образом (см.рис.4):

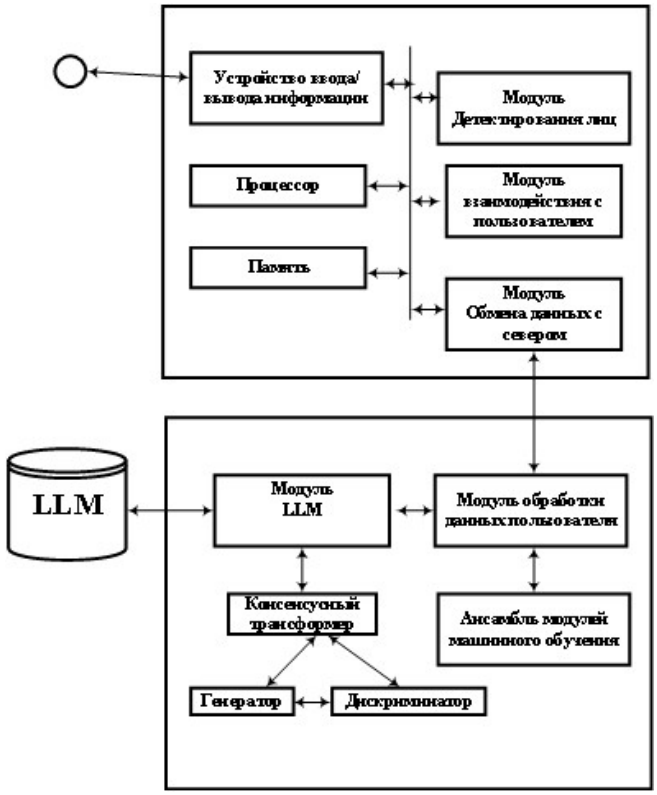

Рис.4. Консенсусный трансформер медицинского ассистента

Предложенное техническое решение обеспечивает решение технической проблемы в части создания более эффективного цифрового интерактивного медицинского ассистента для обеспечения консультаций пользователя. Техническим результатом является расширение функциональных возможностей при формировании информации о

здоровье пользователя за счет обеспечения интерактивного диалога с большой языковой моделью, обогащающей ответы на основании информации о **состоянии** пользователя, получаемой в реальном времени

ЛИТЕРАТУРА

1. Гегель, Г. В. Ф. Наука логики. - Санкт-Петербург: Наука, 2005. - 799 с.
2. Моисеев Н. Н. Человек, природа и будущее цивилизации: "Ядер. зима" и пробл. "запрет. черты". – М.: Изд-во агентства печати "Новости", 1986. – 92 с.
3. Павловский Ю. Н. Метод имитационных игр в проблемах геополитики, безопасности, межгосударственных отношений. // В кн. Материалы учредит. конф. Российского общества исследования операций. - М.: ВЦ РАН. 1997. С. 44-56.
4. Ерешко Ф. И., Белотелов Н. В., Бродский Ю. И., Турко Н. И. Имитационные игры как инструмент исследования геополитических проблем / // Управление развитием крупномасштабных систем (MLSD'2023) : Труды Шестнадцатой международной конференции, Москва, 26–28 сентября 2023 года. – Москва: Институт проблем управления им. В.А. Трапезникова РАН, 2023. – С. 488-494.
5. Ерешко Ф.И.Формализация механизмов принятия решений на примере имитационных игр. Доклад. // Управление развитием крупномасштабных систем (MLSD'2025): Труды Восемнадцатой международной конференции, Москва, 24–26 сентября 2025 года. – Москва: Институт проблем управления им. В. А. Трапезникова РАН, 2025.
6. Воронцов К. В. «Искусственный интеллект: эволюция идей от Фрэнсиса Бэкона до векторных трансформеров и ChatGPT». -М.: МГУ. Институт искусственного интеллекта. // Проблемы искусственного интеллекта – Совместный научный семинар Российской ассоциации искусственного интеллекта и ФИЦ “Информатика и управление” РАН 19 апреля 2023 г.
7. Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (Декабрь 2017). Attention is All you Need. In I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett (ed.). *31st Conference on Neural Information Processing Systems (NIPS)*. Advances in Neural Information Processing Systems. Vol. 30. Curran Associates, Inc. arXiv:1706.03762.
8. Прошина, М. В. Эволюция языковых моделей. // Инновации и инвестиции. – 2023. – № 10. – С. 230-235.
9. Hayes B. First links in the Markov chain. // American Scientist. - 2013. – Vol. 101 (2). – P. 92.-97.
10. Shannon, C. A Mathematical Theory of Communication. // The Bell System Technical Journal. – 1948. – Vol. 27. – P. 379–423.
11. Chomsky, N. Three models for the description of language / N. Chomsky // IEEE Transactions on Information Theory. - 1956. Vol. 2 (3). - P. 113–124.
12. Ерешко Ф.И. Горелов М.А. Игровое представление искусственных нейронных сетей // Анализ, моделирование, управление, развитие социально-экономических систем: сборник научных трудов XVI Международной школы-симпозиума АМУР-2022, Симферополь-Судак, 14-27 сентября 2022 / ред. совет: А. В. Сигал (предс.) и др. – Симферополь : ИП Корниенко А. А., 2022. – 408 с. ISBN 978-5-6043882-9-7. С. 146–149.
13. Korzybski A. Une carte n’est pas le territoire. – P.: Editions de l’Eclat, 1998. – 191 с.
14. Иванова С.А., Суетин А.Г. Работа с информацией: сдвиг парадигмы // Проектирование будущего. Проблемы цифровой реальности: труды 1-й Международной конференции (8-9 февраля 2018 г., Москва). — М.: ИПМ им. М.В.Келдыша, 2018. — С. 152-157.
15. Athul Paul Jacob, Yikang Shen, Gabriele Farina, Jacob Andreas. The Consensus Game: Language Model Generation via Equilibrium Search. Published as a conference paper at ICLR 2024. arXiv:2310.09139v1 . https://doi.org/10.48550/arXiv.2310.09139.
16. Athul Paul Jacob, David J. Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. Modeling strong and human-like gameplay with kl-regularized search. In International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.
17. Теория игр может сделать ИИ более корректным и эффективным. // https://habr.com/ru/companies/first/articles/837292/
18. Воробьев Н.Н. Основы теории игр. Бескоалиционные игры. – М.: Наука. Главная редакция физико-математической литературы. 1984, - 496 с.
19. Блинов П. Д., Егоров К. С., Ботман С. А., Свиридов И. А., Мифтахова А. Р., Зубкова Г. В., Савченко А. В., Кудин С. С. Интеллектуальный медицинский ассистент и способ предоставления пользователю информации о состоянии здоровья с его использованием. / патентообладатель публичное акционерное общество " Сбербанк России" (ПАО Сбербанк) / Российская Федерация. Федеральная служба по интеллектуальной собственности. RU 2 850 170, Заявка: 2024137485, 12.12.2024.

References in Cyrillics

1. Hegel, G. W. F. The Science of Logic. - St. Petersburg: Nauka, 2005. - 799 p.
2. Moiseev, N. N. Man, Nature, and the Future of Civilization: "Nuclear Winter" and the Problem of "Banning the Line." - Moscow: Novosti Press Agency, 1986. - 92 p.
3. Pavlovsky, Yu. N. The Method of Simulation Games in Geopolitics, Security, and Interstate Relations. // In the book. Materials of the founding conf. of the Russian Society of Operations Research. - M.: VTs RAS. 1997. P. 44-56.

4. Ereshko F. I., Belotelov N. V., Brodsky Yu. I., Turko N. I. Simulation Games as a Tool for Researching Geopolitical Problems / // Management of Large-Scale System Development (MLSD'2023) : Proceedings of the Sixteenth International Conference, Moscow, September 26–28, 2023. – Moscow: V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, 2023. – Pp. 488-494.
5. Ereshko F.I. Formalization of Decision-Making Mechanisms on the Example of Simulation Games. Report. // Management of Large-Scale System Development (MLSD'2025): Proceedings of the Eighteenth International Conference, Moscow: V. A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 2025.
6. Vorontsov K. V. "Artificial intelligence: the evolution of ideas from Francis Bacon to vector transformers and ChatGPT." Moscow: MSU. Institute of Artificial Intelligence. // Problems of artificial Intelligence – A joint scientific seminar of the Russian Association of Artificial Intelligence and the Scientific Research Center “Informatics and Management” of the Russian Academy of Sciences on April 19, 2023.
7. Proshina, M. V. Evolution of language models. // Innovation and investment. – 2023. – No. 10. – Pp. 230-235.
8. Ereshko F.I. Gorelov M.A. Game Representation of Artificial Neural Networks // Analysis, Modeling, Management, and Development of Socio-Economic Systems: Collection of Scientific Papers of the XVI International School-Symposium AMUR-2022, Simferopol-Sudak, September 14-27, 2022 / ed. Council: A. V. Sigal (Chairman) et al. – Simferopol: IP Kornienko A. A., 2022. – 408 p. ISBN 978-5-6043882-9-7. Pp. 146–149.
9. Ivanova S.A., Suetin A.G. Working with Information: A Paradigm Shift // Designing the Future. Problems of Digital Reality: Proceedings of the 1st International Conference (February 8-9, 2018), Moscow). Moscow: IPM named after M. N.V. N. By the Way, 2018. pp. 152-157.
10. Game theory can make AI more correct and effective. // https://habr.com/ru/companies/first/articles/837292/
11. Vorobyov N.N. Fundamentals of Game Theory. Non-Cooperative Games. – M.: Nauka. Main Editorial Board of Physical and Mathematical Literature. 1984, - 496 p.
12. Blinov P. D., Egorov K. S., Botman S. A., Sviridov I. A., Miftakhova A. R., Zubkova G. V., Savchenko A. V., Kudin S. S. Intelligent medical assistant and a method for providing a user with information about the state of health using it. / patent holder Public Joint Stock Company "Sberbank of Russia" (PAO Sberbank) / Russian Federation. Federal Service for Intellectual Property. RU 2 850 170, Application: 2024137485, 12.12.2024.

Ерешко Феликс Иванович

доктор технических наук, профессор, зав. Отделом Вычислительный центр им. А. А. Дородницына Российской академии наук Федерального исследовательского центра «Информатика и управление» РАН, 119333, Москва, ул. Вавилова, 40
ORCID: 0000-0002-1732-2204,
fereshko@yandex.ru

Сараев Виктор Никифорович

Кандидат технических наук, Лауреат Премии Правительства РФ в области науки и техники, Первый заместитель генерального директора, Международный научно-исследовательский института, сооснователь АО “РК сложные системы” Фонда Росконгресс
(117312 Россия, Москва, проспект 60-летия Октября, д. 9);
ORCID: 0009-0002-0370-7960
ourtokentrust@mail.ru

ТКАЧЕНКО Аделина Владимировна
Аспирант, Институт проблем управления им. В. А. Трапезникова РАН
(117342, г. Москва, ул. Профсоюзная, д. 65, стр. 2)
ORCID: 0009-0008-4366-5500
adelina.tkch@gmail.com

Ключевые слова: искусственный интеллект, языковые модели, консенсусный трансформер, теория игр, имитация человеческого интеллекта, цифровой интерактивный медицинский ассистент.

Felix I. Ereshko, Viktor N. Saraev, Adelina V. Tkachenko. On the Consensus of Language Models of Artificial Intelligence

Keywords: Artificial intelligence, language models, consensus transformer, game theory, imitation of human intelligence, digital interactive medical assistant.
DOI:
JEL classification: C65-Разнообразные математические инструменты; C71 Кооперативные игры

Abstract

Language models and artificial intelligence transformer models coupled with game theory procedures are considered. In the proposed Scheme for solving specific problems using game-based artificial intelligence, a consensus transformer is designed to imitate the actions of a natural person. The transformer formulates a strategy in natural language.